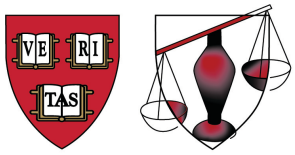


Berkman
The Berkman Center for Internet & Society
at Harvard University

Research Publication No. 2014-12
July 22, 2014

Integrating Approaches to Privacy across the Research Lifecycle: Long-term Longitudinal Studies

Alexandra Wood
David R. O'Brien
Micah Altman
Alan F. Karr
Urs Gasser
Michael Bar-Sinai
Kobbi Nissim
Jonathan Ullman
Salil Vadhan
Michael Wojcik

23 Everett Street • Second Floor • Cambridge, Massachusetts 02138
+1 617.495.7547 • +1 617.495.7641 (fax) • http://cyber.law.harvard.edu •
cyber@law.harvard.edu

Integrating Approaches to Privacy across the Research Lifecycle: Long-term Longitudinal Studies

Alexandra Wood, David R. O'Brien; Micah Altman, Alan F. Karr, Urs Gasser, Michael Bar-Sinai, Kobbi Nissim, Jonathan Ullman, Salil Vadhan, Michael Wojcik[1]

WORKING PAPER

[1] Alexandra Wood and David R. O'Brien were the lead authors, having written the initial draft of the manuscript and taken primary responsibility for revisions. The workshop committee (Micah Altman, Michael Bar-Sinai, Kobbi Nissim, David R. O'Brien, Jonathan Ullman, Salil Vadhan, Michael Wojcik, and Alexandra Wood) contributed to the conception of the report (including core ideas and statement of research questions), to the methodology (development of the workshop questions, use cases, conceptual model, and taxonomies applied), to the project administration (coordination and management of the workshop and report writing process), and to the writing through critical review and commentary. Alan F. Karr and Urs Gasser contributed to report conception, by providing and articulating ideas as facilitators of the workshop session. Micah Altman, as workshop committee chair, and Salil Vadhan supervised the workshop process and report writing. Salil Vadhan, as managing PI of the Privacy Tools for Sharing Research Data project, led the funding acquisition supporting this workshop. The workshop and the writing of this report were supported by NSF grant CNS-1237235.

The co-authors would like to thank their fellow workshop participants for their contributions to the workshop and this report: John Abowd, Edoardo Airoldi, Aslan Askarov, Boaz Barak, Khaliah Barnes, Raef Bassily, Kristen Bolt, Scott Bradner, Mark Bun, Ran Canetti, Kenneth L. Carson, Eleni Castro, Stephen Chong, Mercè Crosas, Cynthia Dwork, Robert Gellman, Sharon Goldberg, Raquel Hill, Murat Kantarcioglu, Peter Katz, Henry Lam, David Lazer, Wendy Mariner, Dan O'Brien, Davide Proserpio, Sofya Raskhodnikova, Leonid Reyzin, Aleksandra Slavkovic, Adam Smith, Greta Lee Splansky, Peter Suber, Latanya Sweeney, Aurelia Tamò, Adam Tanner, Kit Walsh, John Wilbanks, Christopher Winship, Felix Wu, and Tanya Zlateva.

1 Background

The science of understanding human behavior, health, and interactions is being transformed by the ability of researchers to collect, analyze, and share data about individuals on a wide scale. However, a major challenge for realizing the full potential of such data science is ensuring the privacy of human subjects. And as new demonstrations and methods of reidentification continue to emerge, traditional approaches to protecting privacy are becoming decreasingly effective.

On September 24-25, 2013, the Privacy Tools for Sharing Research Data project at Harvard University, in collaboration with the Reliable Information Systems and Cyber Security Center at Boston University, held a workshop titled "Integrating Approaches to Privacy across the Research Lifecycle." Over forty leading experts in computer science, statistics, social science, law, and policy convened to discuss the state of the art in data privacy research. Participants considered how emerging tools and approaches from their various disciplines should be integrated in the context of real-world use cases involving the management of confidential research data.

This paper is part of a larger body of workshop materials summarizing the tools and use cases discussed and mapping out, at a high level, a research agenda to advance the integration of various methods of preserving confidentiality in research data. Additional materials produced in conjunction with the workshop are available on the workshop website.[2]

Introduction to the data privacy use cases

During the workshop, participants discussed three use cases, each involving a different category of data: computational social science data, access to journal publication data, and long-term longitudinal data. In addition, an official statistics use case was identified as a base case for comparison against the three use cases discussed during the workshop. These use cases were designed to raise different questions and challenges related to the management of confidential data.

The computational social science use case describes the advancement of statistical and data analytic methods and the integration of vast amounts of data from multiple sources in the study of humans, human behavior, and human institutions. Cutting edge social science research now incorporates, for instance, the analysis of social network data, harvested textual data, professional and amateur video content, and fine-grained geospatial data including the traces of movements in space and time. Examples of computational social science data include Netflix movie viewing records,[3] GPS location data,[4] blog postings,[5] Facebook data,[6] and data from urban sensor

[2] Privacy Tools for Sharing Research Data, "Integrating Approaches to Privacy across the Research Lifecycle," http://privacytools.seas.harvard.edu/fall-2013-workshop.
[3] *See* Arvind Narayanan and Vitaly Shmatikov, "Robust De-anonymization of Large Sparse Datasets," *Proceedings of the 2008 IEEE Symposium on Security and Privacy* (2008).
[4] *See* Dale L. Zimmerman and Claire Pavlik, "Quantifying the Effects of Mask Metadata, Disclosure, and Multiple Releases on the Confidentiality of Geographically Masked Health Data," *Geographical Analysis* 40 (2008): 52-76.

systems.[7]

The next use case, access to journal publication data, involves the publication of scientific research in peer-reviewed journals and the related principles of reproducibility, replicability, and extensibility. Many of the findings published in such journals, especially those in the medical, health, behavioral, and social sciences, do not describe methods in sufficient detail to support verification and replication. At the same time, pressure from various sources to make data more openly available has increased. Examples of journal policies for data sharing and citation include the data sharing systems of open access journals[8] and the American Political Science Association Data Access and Research Transparency Policy Initiative.[9]

The third use case addresses long-term longitudinal data, or data collected from a set of human subjects at multiple points over a long period of time. Longitudinal data may include a wide variety of highly-specific and often sensitive information about individuals, such as information related to their health, socioeconomic, and behavioral characteristics, and such data are often made available to researchers in microdata form as public-use or restricted-use datasets.[10] Examples of longitudinal studies include the Framingham Heart Study[11] and the Panel Study of Income Dynamics.[12]

Finally, the use case underlying the workshop discussions describes official statistics, or statistics produced by government agencies using surveys, censuses, and other forms of data collection. The data are typically disseminated to the public as aggregate data or sanitized microdata to support policy analysis, public transparency, and research purposes. Examples of official statistics include the US Census[13] and European national statistical offices.[14]

---

[5] *See* Ahmed Al Faresi et al., "Privacy Leakage in Health Social Networks," *Computational Intelligence* (2013), doi:10.1111/coin.12005.

[6] *See* Carter Jernigan and Behram F.T. Mistree, "Gaydar: Facebook Friendships Expose Sexual Orientation," *First Monday* 14, no. 10 (2009); Lars Backstrom et al., "Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography," *Proceedings of the 16th International World Wide Web Conference* (2007); Michael Zimmer, "But the Data Is Already Public: On the Ethics of Research in Facebook," *Ethics and Information Technology* 12, no. 4 (2010): 313-325.

[7] *See, e.g.,* Boston Area Research Initiative, http://www.bostonarearesearchinitiative.net.

[8] *See, e.g.*, "PKP-Dataverse Integration Project: Dataverse Network Collaboration with the Public Knowledge Project," last visited Nov. 6, 2013, http://projects.iq.harvard.edu/ojs-dvn.

[9] *See* American Political Science Association, *Draft Guidelines for Data Access and Research Transparency for Qualitative Research in Political Science* (August 7, 2013).

[10] Microdata are "data on the characteristics of units of a population, such as individuals, households, or establishments, collected by a census, survey, or experiment." US Census Bureau, *Survey Design and Statistical Methodology Metadata*, IT Standard 16.0.0 (Aug. 1998), http://www.census.gov/srd/www/metadata/metada18.pdf.

[11] "Framingham Heart Study," last updated September 11, 2013, http://www.framinghamheartstudy.org.

[12] "Panel Study of Income Dynamics," last visited October 21, 2013, http://psidonline.isr.umich.edu.

[13] "United States Census Bureau," last visited October 21, 2013, http://www.census.gov; see also United States Census Bureau, *Census Confidentiality and Privacy, 1790-2002* (2004), http://www.census.gov/prod/2003pubs/conmono2.pdf.

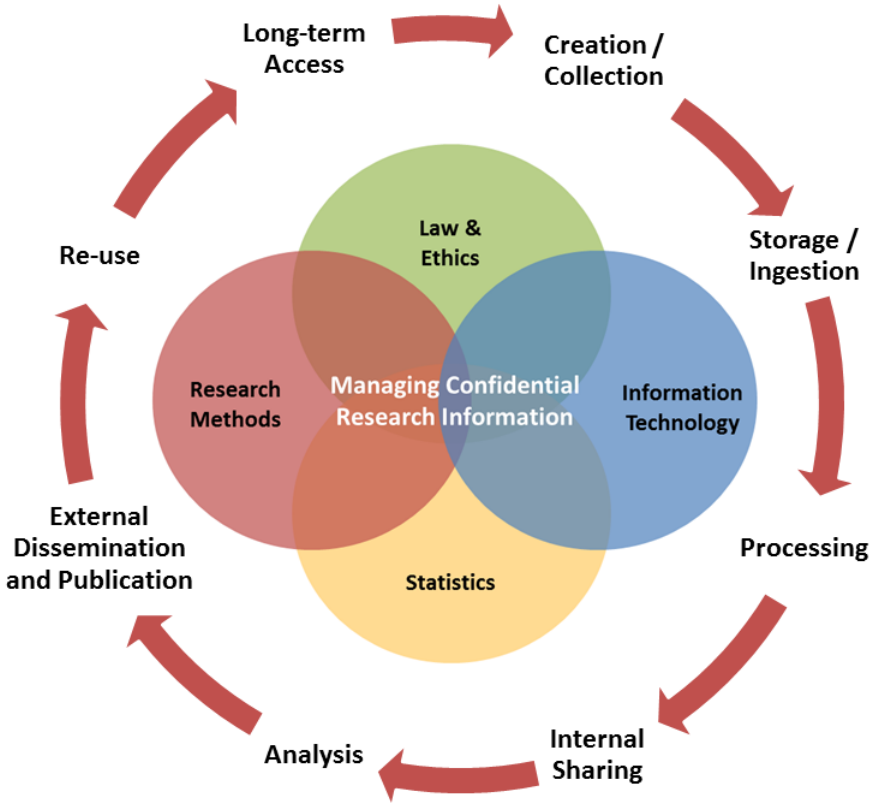
[14] *See* A Network of Excellence in the European Statistical System in the field of Statistical Disclosure Control (ESSNET), *Handbook on Statistical Disclosure Control* (January 2010), http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf.

Framework for analysis

In exploring the three data privacy use cases, workshop participants considered the characteristics that are most important to selecting data management approaches. They also examined the tradeoffs between privacy and utility and the barriers and incentives associated with protecting confidential information. The discussions were guided by framing materials shared in advance of the workshop, including an overview of the research information lifecycle, a chart of example use case features, and a list of discussion questions.

The research information lifecycle, illustrated below in Figure 1, depicts the common lifecycle stages from the creation and collection stage through the long-term access stage.[15] In the center of the lifecycle, there is an illustration of four categories of approaches to managing confidential research information—research methods, information technology, statistics, and law and ethics—and the intersections between them.

Figure 1. Example Research Information Lifecycle



[15] *See* Sarah Higgins, "The DCC Curation Lifecycle Model," *International Journal of Digital Curation* 3, no. 1 (2008): 134-140, http://www.dcc.ac.uk/resources/curation-lifecycle-model.

A chart of example use case features, presented below in Table 1, summarizes several groups of features that have previously been used to characterize these and similar use cases.[16] This typology of use case characteristics is intended to abstract out their key features. This chart was provided in order to focus discussions of the three data privacy use cases, identify the sources of privacy and management challenges, and delimit the applicability of proposed approaches to data and confidentiality management.

Table 1. Example Use Case Features

Families of Features	***Specific Characteristics***
Data characteristics	• Logical Structure (e.g., single relation, multiple relational, network/graph, semi-structured, geospatial, aggregate table) • Source • Unit of observation • Attribute measurement type (e.g., continuous/discrete; ratio/interval/ordinal/nominal scale; associated schema/ontology) • Performance characteristics (e.g., dimensionality/number of measures, number of observation/volume, sparseness, heterogeneity/variety, frequency of updates/velocity) • Quality characteristics (e.g., measurement error, metadata, completeness, total error)
Disclosure scenarios	• Source of threat (e.g., natural, unintentional, intentional) • Areas of vulnerability (e.g., data, software, logistical, physical, social engineering) • Attacker objectives, background knowledge, and capability (e.g., "nosy neighbor," "business competitor," "muckraking journalist," "panopticon," "intrusive employer/insurer") • Breach criteria/disclosure concept
Legal/institutional context of data collection	• Consent (e.g., open consent, active but limited consent, passive/implicit consent, awareness of data collection, unawareness of data collection, surreptitious data collection) • Jurisdiction where collection takes place • Special legal relationship with subject (e.g., student relationship under FERPA, patient relationship under HIPAA) • Status of individual/institution responsible for data collection (e.g., a HIPAA regulated entity, an entity subject to the Common Rule, 45 C.F.R. part 46)

[16] *See, e.g.,* Leon Willenborg and Ton de Waal, *Elements of Statistical Disclosure Control*, vol. 155, *Lecture Notes in Statistics* (New York: Springer Verlag, 2001); A Network of Excellence in the European Statistical System in the Field of Statistical Disclosure Control (ESSNet SDC), *Handbook on Statistical Disclosure Control* (January 2010), http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf; Benjamin Fung, et al., "Privacy-preserving Data Publishing: A Survey of Recent Developments," *ACM Computing Surveys* (CSUR) 42, no. 4 (2010): 14; Micah Altman, "Mitigating Threats to Data Quality Throughout the Curation Lifecycle" (position paper from a workshop, Curating For Quality: Ensuring Data Quality to Enable New Science, Arlington, Virginia, September 10-11, 2012), http://datacuration.web.unc.edu.

Information lifecycle stages	• Lifecycle stages managed/in scope (e.g., data creation and experimental intervention; data collection and initial transmission; data storage, ingestion, entry into research environment; processing; internal sharing and collaboration; analysis; dissemination; publication; verification, scientometric, educational, and scientific reuse; long-term access) • Information management policies
Analytic results	• Form of output (e.g., summary scalars, summary table, model parameters, data extract, static data publication, static visualization, dynamic visualization, statistical/model diagnostics) • Analysis methodology (e.g., contingency tables/counting queries, summary statistics/function estimation, regression models/GLM, general model-based statistical estimation/MLE/MCMC, bootstraps/randomization/data partitioning, data mining/heuristics/custom algorithms) • Analysis goal (e.g., rule-based, theory formation, existence proof, verification, descriptive inference, forecasting, causal inference, mechanistic inference) • Utility/loss/quality measure (e.g., entropy, mean squared error, realism, validity of descriptive/predictive/causal statistical inference)
Characteristics of data related to informational harm	• Attribute identifiability characteristics (e.g., direct identifiers, quasi-identifiers/publicly observable fixed characteristics of individuals) • Attribute sensitivity (e.g., descriptions of criminal conduct, health status, income, political affiliations) • Statistical identifiability risks (e.g., reidentification/record-linkage risks, information/learning risks) • Expected types of harms from reidentification (e.g., loss of insurability, loss of employability, market discrimination, criminal liability, psychological harm, loss of reputation, emotional harm, and loss of dignity (dignitary harm); social harms to a vulnerable group (e.g., stereotyping), price discrimination against vulnerable groups, market failures; chilling of speech and action; potential for political discrimination; potential blackmail and other abuses) • Expected magnitude of harm, if identification occurs (e.g., minimal, moderate, severe)
Lifecycle stakeholders	• Stakeholder types (e.g., consumer, producer, funder, host institution, researcher, regulator, subject, citizen, journal) • Stakeholder capacities/resources (e.g., technical expertise, infrastructural capacity, budget, staffing resources) • Trust relationships
Current approaches	• Regulations/policy • Legal controls • Statistical/computational disclosure control methodology • Information security controls

The following is a list of discussion questions provided to participants to guide their conversations as they analyzed the data privacy use cases. These questions were designed to encourage the participants to consider the defining features of the use cases, the state of the art of current approaches to data confidentiality, the tradeoffs between privacy and utility, and ways to enhance and integrate current approaches to data privacy.

- *Characterization*. Are there key additional characteristics of the use cases that should be noted? How do these characteristics change the analysis and treatment of privacy in these use cases?
- *Current approaches.* How is confidentiality in the use cases currently managed? What is the state of the art and practice? How is success measured?
- *Enhancing approaches*. Are any of the legal and technical approaches discussed during the first day of the workshop used in practice? How could the tools and approaches mentioned earlier (or other existing tools) be used at particular stages of the research lifecycle to enhance privacy and utility?
- *Integrating approaches.* Are approaches that have been developed and used in different communities compatible with one another? How should legal, computational, policy, and statistical tools be integrated to be most effective?
- *Utility*. What would stakeholders like to do with the data that the toolset does not restrict or obstruct? Where is the social benefit sub-optimal? How is utility measured or perceived by the stakeholders?
- *Privacy*. What types of data and outputs are considered particularly sensitive? What are the most important real and perceived risks? What harms could occur if data are released and reidentified? How severe are these harms, and how likely are they to occur?
- *Methodological Barriers.* What are the technical, methodological, computational or infrastructural barriers to improving privacy and utility in the management of these data? What particular characteristics of the use case contribute to such barriers?
- *Incentives*. If better tools already exist, why are they not being used? What are the barriers to adopting new tools and methods? What are the specific market failures in this area, such as perverse incentives, lack of or asymmetry of information, lack of a well-developed market, irrational behavior, transaction costs, or network effects? What particular characteristics of the use case most influence incentives?
- *Future*. How is this use case likely to evolve over time? What are the threats to stability, scalability, robustness, and resilience of the current and proposed solutions?
- *Prior work.* Are there key additional examples of the use case that should be noted? Are there additional key references or writings that should be noted?

2 Introduction to long-term longitudinal data

The goal of this paper is to analyze the long-term longitudinal data use case. It defines the features characteristic to long-term longitudinal studies, describes challenges that arise from their defining characteristics, identifies risks associated with managing confidentiality in longitudinal data, examines common approaches currently used to protect the confidentiality of the data, and outlines urgent problems and areas for future research on these topics.

What is a longitudinal study?

A longitudinal study collects information from the same respondents at multiple points over time. Longitudinal studies are often conducted over an extended period of time, though timeframes may vary widely. This category of study may include both a study conducted over a four-year period[17] and a study conducted across many decades.[18] In addition, the data may be collected at frequent intervals, such as monthly or quarterly, or less frequently, such as every one, five, or ten years.

The types of information collected, as well as the purposes and contexts for such collection, can vary substantially among longitudinal studies. For instance, researchers often collect longitudinal data from respondents in economic, behavioral, and epidemiological studies, and link survey and administrative data to perform policy analyses. Collection methods include interviews, self-completion questionnaires, direct measurements, blood and tissue samples, and records requests, and the unit of analysis may be the individual, a household or a parent-child group, or a workplace, school, or other establishment.

In certain contexts, longitudinal data may offer advantages over cross-sectional data, or data collected at a single point in time. Because longitudinal data allow researchers to measure changes at the individual level over time, the data enable researchers to study trends and events throughout individual lifetimes or even generations. Researchers may also use longitudinal data to analyze and compare differences among individual patterns of change over time.[19]

Exemplar long-term longitudinal datasets

This paper focuses, in particular, on longitudinal studies that collect data about individuals over a long period of time. Some of the exemplar datasets identified during the workshop include the Framingham Heart Study,[20] the Survey of Doctorate Recipients,[21] Project Talent,[22] the Panel

[17] *See, e.g.,* Andrew T. A. Cheng, et al., "A 4-Year Longitudinal Study on Risk Factors for Alcoholism," *Archives of General Psychiatry* 61, no. 184 (2004).
[18] *See, e.g.*, "Framingham Heart Study," last updated September 11, 2013, http://www.framinghamheartstudy.org.
[19] *See generally* Judith D. Singer and John B. Willett, "A Framework for Investigating Change over Time," chap. 1 in *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence* (New York: Oxford University Press, 2003).
[20] "Framingham Heart Study," last updated September 11, 2013, http://www.framinghamheartstudy.org/index.html.

Study of Income Dynamics,[23] the Health and Retirement Study,[24] the National Longitudinal Surveys,[25] and the National Education Longitudinal Study of 1988.[26] The following descriptions of these datasets provide an overview of some of the types of data collected, methodologies and timeframes utilized, populations surveyed, and attributes and relationships studied in long-term longitudinal studies.

- The **Framingham Heart Study** (FHS) is a long-term, ongoing cardiovascular study that began in Framingham, Massachusetts, in 1948 with a population of 5,209 men and women between the ages of 30 and 62. Every two years, the FHS, under the direction of the National Heart, Lung and Blood Institute, conducts physical and psychological examinations of and performs laboratory tests on blood and tissue samples from study participants. Since the start of the study, the FHS has studied three generations of participants, resulting in the collection of health information from nearly 15,000 individuals.

- The **Survey of Doctorate Recipients** (SDR) is conducted by the National Opinion Research Center at the University of Chicago for the National Science Foundation and the National Institutes of Health. The study focuses on individuals who received a doctoral degree in science, engineering, or health from an academic institution in the United States. Data collected through the SDR are related to the education, training, work experience, and career development of this population.

- **Project Talent** is a national longitudinal study begun in 1960 with a population of 440,000 students in grades nine through twelve from 1,353 schools. Researchers administered subject tests that evaluated participants' competencies in mathematics, science, and reading comprehension. They also distributed questionnaires covering family background, personal and educational experiences, aspirations for future education and vocation, and interests in various occupations and activities. Participants were re-contacted three times—at one, five, and eleven years after high school graduation—with questionnaires on educational and occupational attainment, marriage, family formation, and other topics. Efforts to re-contact participants and collect additional data are currently underway.

---

[21] The National Opinion Research Center at the University of Chicago, "Survey of Doctorate Recipients," last visited October 21, 2013, https://sdr.norc.org.
[22] American Institutes for Research, "Project Talent," last visited October 21, 2013, http://www.projecttalent.org.
[23] University of Michigan, "PSID: A National Study of Socioeconomics and Health Over Lifetimes and Across Generations," last visited October 21, 2013, http://psidonline.isr.umich.edu.
[24] University of Michigan, "Health and Retirement Study: A Longitudinal Study of Health, Retirement, and Aging Sponsored by the National Institute on Aging," last visited October 21, 2013, http://hrsonline.isr.umich.edu.
[25] Bureau of Labor Statistics, "National Longitudinal Surveys," last visited October 21, 2013, http://www.bls.gov/nls.
[26] National Center for Education Statistics, "National Education Longitudinal Study of 1988," last visited October 21, 2013, http://nces.ed.gov/surveys/nels88.

- The **Panel Study of Income Dynamics**, begun in 1968 by the University of Michigan, studies a nationally representative sample of over 18,000 individuals living in 5,000 families in the United States. Researchers have collected socioeconomic and health information on these individuals and their descendents continuously. Topics included in this dataset are employment, income, wealth, expenditures, health, marriage, childbearing, child development, philanthropy, and education.

- The **Health and Retirement Study** is a longitudinal panel study initiated by the University of Michigan in 1992. Every two years, researchers collect information from a representative sample of more than 26,000 Americans over the age of 50. They use interviews to collect data on income, work, assets, pension plans, health insurance, disability, physical health and functioning, cognitive functioning, and health care expenditures. These data are used to study the changes in labor force participation and the health transitions that individuals undergo toward the ends of their work lives and into retirement.

- The **National Longitudinal Surveys** (NLS) are a set of surveys designed by the Bureau of Labor Statistics to gather information on the labor market activities and other significant life events of several groups of men and women. The National Longitudinal Survey of Youth 1997 is an annual survey begun in 1997 of young men and women born in the years 1980-84. The National Longitudinal Survey of Youth 1979 (NLSY79) is an annual survey of a nationally representative sample of 12,686 young men and women who were 14-22 years old when they were first interviewed in 1979. Data from the children of the women in this study were collected beginning in 1986 in the NLSY79 Children and Young Adults survey. The NLS also included national longitudinal studies of cohorts of young men and women and older men and women.

- The **National Education Longitudinal Study of 1988**, one of many longitudinal studies conducted by the National Center for Education Statistics, is a study of a nationally representative sample of individuals who were in eighth grade in 1988. Participants were re-contacted in 1990, 1992, 1994, and 2000. Questionnaires distributed to students covered topics such as school, work, and home experiences; educational resources and support; the role of their parents and peers in their education; neighborhood characteristics; educational and occupational aspirations; and self-reports on smoking, alcohol and drug use and extracurricular activities. In addition, the study administered achievement tests in reading, social studies, mathematics, and science; surveyed the students' teachers, parents, and school administrators; and collected coursework and grades from students' high school as well as their postsecondary transcripts.

These seven exemplar datasets served as a foundation for the analysis of long-term longitudinal studies undertaken during the workshop. They are also referenced throughout this paper's discussions of current approaches to collecting and maintaining confidential research data.

Topics covered during long-term longitudinal study use case discussions

Participants discussing the long-term longitudinal study use case centered their analysis around three principal questions:

- What privacy and utility issues are unique to longitudinal data?
- What are current best practices?
- What are the most urgent unsolved problems?

The resulting conversations covered a wide range of topics and informed participants' understanding of the defining features of longitudinal studies and related challenges; the risks associated with longitudinal data; common practices related to the collection, storage, use, and sharing of longitudinal data; and areas where further research is needed.

3 Defining features of longitudinal studies and related challenges

The following is a discussion of the defining features of longitudinal studies, as well as some of the challenges that researchers may encounter as a result of these features. This section utilizes the research information lifecycle model, illustrated in Figure 1 above, as a framework for analysis, with particular emphasis on the creation and collection, storage and ingestion, processing, analysis, external dissemination, and reuse stages.

At the data creation and collection stage

Defining features of long-term longitudinal studies at the data creation and collection stage include the extended timeframe for collection, the laws and regulations that govern data collection in certain sectors, the highly-specific and often very sensitive nature of the data collected, and the role of institutions and institutional review boards in designing and conducting a research study. These features and the challenges they present are described in detail below.

Long-term longitudinal studies require researchers to collect data from the same set of subjects over a long period of time. This raises a number of issues at the data creation and collection stage because the types of data collected, the ways the data are used, the analytical methods, the methods for identifying privacy risks, and privacy regulations are all likely to change over the lifetime of the study. The longer the study timeframe, the greater the likelihood that these circumstances will change after the initial data collection, and the more likely it is that issues will arise related to participant consent and compliance with privacy laws and regulations.

As a threshold matter, federal and state laws and regulations and institutional policies often require researchers to obtain informed consent from participants before a study can commence.[27] Although the main objective is to facilitate a voluntary decision to participate, the informed consent process often serves multiple purposes. For instance, it functions as a communication channel where the participant is informed of the nature and scope of the study, the type of

[27] *See, e.g.*, 45 C.F.R. §§ 46.116-117.

information sought, the extent to which confidentiality will be preserved in the study, and any risks related to participation. Consent takes the form of a written, legally enforceable contract that is preserved for documentation purposes by the researchers.

When a study is conducted over a long period of time, the categories of information collected may evolve as study investigators or funders change or new substantive research questions and hypotheses emerge. The information collected from a research subject may also change over time for other reasons. For example, a study may initially collect data on high school students' academic outcomes and later transition to collecting data about the respondents' labor market participation, marriage status, and family size.[28] Study investigators face the challenge of sufficiently describing in the consent form the categories of information to be collected over the course of the study and, if the scope of the information to be collected expands over time, obtaining new consent from the respondents.[29] Consent issues may also arise when the timeframe of a study is extended beyond the timeframe initially disclosed to the study participants. In addition, changes in the population being studied may create challenges related to consent. For example, participants may decide to withdraw their participation in a study for a variety of reasons. Long-term studies trace patterns across generations by including the children and grandchildren of the original respondent as participants. In these studies, parents may provide the initial consent for a child, and questions regarding consent may arise when the child reaches the age of majority and becomes able to provide informed consent on his or her own behalf.

The types of personal information collected in a longitudinal study may be highly-specific and are often very sensitive. Longitudinal studies sometimes collect data about respondents' income and assets, sexual activity, criminal activities, and drug and alcohol use.[30] The highly sensitive nature of such information influences the study administrators' selection of data collection methods, such as the choice between interviews or self-administered questionnaires,[31] as well as their choice of data confidentiality and security procedures, which may offer varying degrees of protection.

Certain types of data collected over the course of a longitudinal study may be protected by federal or state privacy laws, and the rules that apply to any given researcher may vary considerably. For instance, many researchers must comply with the consent requirements for human subjects research outlined in the Common Rule,[32] and some researchers will be subject to the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.[33] Researchers may seek to

---

[28] *See, e.g.,* National Center for Education Statistics, "National Education Longitudinal Study of 1988: Design," last visited October 29, 2013, http://nces.ed.gov/surveys/nels88/design.asp.
[29] *See* National Bioethics Advisory Commission, *Ethical and Policy Issues in Research Involving Human Participants* (2001), https://bioethicsarchive.georgetown.edu/nbac/pubs.html; National Bioethics Advisory Commission, *Research Involving Biological Materials: Ethical issues and Policy Guidance* (2001), https://bioethicsarchive.georgetown.edu/nbac/pubs.html.
[30] *See, e.g.*, Bureau of Labor Statistics, "Confidentiality and Consent," chap. 9 in the *National Longitudinal Survey Handbook* (2005), 153, http://www.bls.gov/nls/handbook/2005/nlshc9.pdf.
[31] *Id.* at 154.
[32] 45 C.F.R. pt. 46.
[33] 42 U.S.C. §§ 1301 et seq.

collect participants' high school and postsecondary transcripts,[34] medical records,[35] or substance abuse treatment records, which may be regulated according to different sets of rules under the Family Educational Rights and Privacy Act (FERPA),[36] HIPAA, or the federal alcohol and drug abuse confidentiality regulations,[37] respectively. Researchers collecting data on behalf of federal agencies may be required to establish certain confidentiality procedures in order to comply with the Privacy Act of 1974[38] or the Confidential Information Protection and Statistical Efficiency Act (CIPSEA).[39] Researchers are tasked with understanding the statutory and regulatory use and access restrictions, data linkage rules, and data security and retention standards that vary widely depending on the categories of information collected and how the data were obtained.

The data collected in longitudinal studies are often rich enough to support analysis methods and research questions not originally envisioned in the study design or in the long-term management of the data. A researcher designing a study cannot anticipate exactly what data collection protocols will be used throughout the life of the study. The original researchers and, to a greater extent, secondary researchers may wish to use data in ways that differ from the original purposes of a study or in ways not anticipated at the time consent was obtained. In these cases, the scope of subjects' consent regarding the use of the data may be unclear or will be limited to the original purposes enumerated in the consent agreement, and new consent may be required. Obtaining new consent is administratively difficult and costly in many cases, and, for researchers conducting analyses long after the original collection, it may not be feasible at all.

Analytical techniques available to researchers, risks of disclosure, and subjects' expectations of privacy may evolve over time. Technological developments have, for example, led to greater capacities for genetic analysis and increased the likelihood that an individual's characteristics can be derived from biological samples collected during a longitudinal study. Techniques for linking study data to external administrative sources, such as government records, have also become more effective.[40] Reidentification risks and best practices for statistical disclosure limitation may change over time, and researchers—not to mention participants—have difficulty judging whether the confidentiality protections used in a study and disclosed in a consent form are adequate. It can be especially difficult to predict disclosure risks for data that may be linked to external datasets or reused in other ways. Along with such technological changes, subjects' expectations of privacy may also shift to some degree, as they become more aware of the increased risk of reidentification or disclosure of information they had not intended to reveal.[41] Accordingly, the

---

[34] *See, e.g.*, *id.*
[35] *See* "Framingham Heart Study," last updated September 11, 2013, http://www.framinghamheartstudy.org/index.html.
[36] 20 U.S.C. § 1232g.
[37] 42 C.F.R. pt. 2.
[38] 5 U.S.C. § 552a.
[39] Pub. L. 107-347, Title V; 44 U.S.C. § 3501 note.
[40] *See, e.g.*, Latanya Sweeney, Akua Abu, and Julia Winn, "Identifying Participants in the Personal Genome Project by Name," Data Privacy Lab, IQSS, Harvard University (2013), http://privacytools.seas.harvard.edu/files/privacytools/files/1021-1.pdf.
[41] *See, e.g.*, Aisling Ann O'Kane, Helena M. Mentis, and Eno Thereska, "Non-Static Nature of Patient Consent: Shifting Privacy Perspectives in Health Information Sharing," *ACM Conference on Computer Supported Cooperative Work* (February 2013), http://research.microsoft.com/apps/pubs/default.aspx?id=179620.

balance between the known benefits and risks of a study may change over time and lead to revisions in the language used in consent forms or, in rare cases, the discontinuation of a study.

Privacy laws and regulations that govern the collection, use, and sharing of personal information may change over time. For example, revisions made in 2011 to FERPA eased restrictions on the sharing of personally identifiable information between state agencies and research organizations for the purpose of improving education programs.[42] Such developments in the law affect the types of information that may be collected, the consent procedures that are required for the collection, and the restrictions that limit use and sharing of the data over the course of the study. However, these changes in the law and how it is interpreted can be unpredictable, which creates uncertainty over the course of a long-term research study.

Privacy and data security regulations and guidelines are not limited to US federal law. There are many relevant privacy laws at the state level, such as state laws protecting the privacy of student records. International issues can also be very complex because many foreign jurisdictions have adopted comprehensive privacy laws that apply to researchers or, in some cases, limit exports of data to countries that do not provide adequate data protection. In addition, researchers may take into account industry best practices and standards when collecting, using, storing, and disclosing sensitive personal information.

Despite the large number of privacy laws and regulations on the books, some researchers will not be subject to any data privacy or security laws. This is especially likely for non-commercial researchers or secondary recipients of records that would otherwise be covered. For example, if a researcher who is not a covered entity receives protected health information from a HIPAA-covered entity, HIPAA's privacy and security rules do not apply to the researcher's secondary use and disclosure of the information.

Researchers will also be required to seek approval from their institutions and institutional review boards (IRBs) prior to commencing a study involving human subjects. Longitudinal studies very often receive federal funding and involve obtaining personal information from or about an individual. Therefore, researchers who conduct longitudinal studies must consider whether the Common Rule[43] governs their research. Whether a study is subject to the IRB process, and its rules related to consent, confidentiality, and dissemination, is determined according to a case-by-case standard, the application of which can be difficult to predict. Regardless of whether the Common Rule applies, longitudinal studies may be subject to the policies of the researcher's institution, which often extend beyond the requirements of the Common Rule. IRB policies may have specific requirements for obtaining informed consent from participants; establishing data collection, use, and dissemination restrictions; and developing and implementing data confidentiality and security procedures for protecting participants' personal information. These institutional requirements play a significant role in a researcher's longitudinal study design, particularly in a researcher's choice of methodologies and confidentiality procedures.

[42] 76 Fed. Reg. 75604-75660.
[43] *See* 45 C.F.R. pt. 46.

Legal and regulatory requirements, institutional policies, and ethical considerations are not the only factors that motivate researchers to establish stringent consent and confidentiality requirements. The defining features and challenges of working with long-term longitudinal data are also related to a researcher's goal of sustaining participant commitment over time. Participant retention, which is strongly associated with data quality and a researcher's ability to continue a study, varies significantly among longitudinal studies, with retention rates ranging anywhere from 19 to 90 percent.[44] Concerns about data confidentiality and security may reduce participation in longitudinal surveys both at the outset of the study and throughout the course of the study. Accordingly, robust consent and confidentiality procedures—and adequate communication about these procedures and the attendant disclosure risks—are key to maintaining participant trust and retention throughout a longitudinal study and are therefore critical to the success of such a study.

At the data ingestion, storage, and processing stages

Several features of longitudinal data create challenges for researchers at the storage and processing stages of the research data lifecycle. These include the need to store study participants' personal identifiers for a long period of time, uncertainty about data ownership and custody rights, and the high institutional costs of managing data storage and access as longitudinal datasets are analyzed, stored, and shared with others.

In order to re-contact subjects for each wave of data collection in a longitudinal study, researchers must retain the subjects' contact information for administrative purposes and be able to link it with identifiers in a dataset over the course of the study. In a long-term longitudinal study collecting data about a large number of individuals spread out across a vast geographic area, many contractors and field investigators may be provided with subjects' personal identifiers for the purpose of contacting them for interviews. Investigators may also need to retain personal identifiers of the subjects' family members and friends in order to track respondents between collection waves. Techniques such as encrypting or hashing identifiers can be used to protect confidentiality when linking longitudinal data, while allowing approved individuals to access the identifiers if necessary. The need to store and share personal identifiers of subjects and their family members and friends creates data confidentiality and security challenges for researchers and field investigators. Researchers' implementation of data security plans and training programs is particularly important in this context.

When it comes to storing longitudinal data, questions of data ownership and custody are sometimes complex or uncertain. There are many players involved in the ingestion and storage of data. Universities, data repositories, funders, principal investigators and other internal

[44] The response rates to the Project Talent follow-up mail surveys ranged from 62 percent for the 12th-grade one-year follow-up to 19 percent for the 9th-grade 11-year follow-up. Project Talent, "Study Design," last visited October 21, 2013, http://www.projecttalent.org/about/studydesign. In contrast, the National Longitudinal Survey of Youth 1979 reports retention rates for respondents between 1979 and 1993 exceeded 90 percent, and between 1994 and 2010 exceeded 75 percent. Bureau of Labor Statistics, "National Longitudinal Survey of Youth 1979: Retention & Reasons for Noninterview," last visited October 25, 2013, https://www.nlsinfo.org/content/cohorts/nlsy79/intro-to-the-sample/retention-reasons-noninterview.

researchers, contractors, and secondary researchers may all wish to claim full or partial ownership of the data at various points throughout the research information lifecycle. Each of these players may have different rights and responsibilities with regards to the data and the human subjects. To address this issue, the parties involved in a transfer of data often require the signing of a data use agreement. Data use agreements governing data transfer and use generally designate one of the parties as the owner of the data and others as licensees permitted to use the data for specific purposes for a limited amount of time.

Also, the cost to institutions of implementing and enforcing privacy safeguards for the storage of longitudinal data is significant. In order to comply with statutory requirements, institutional review board policies, and contractual obligations, institutions are often required to implement comprehensive data security plans, utilize statistical disclosure limitation methods to protect confidentiality in data files, and establish policies and procedures, such as data enclaves, for restricting access to and use of sensitive data. High costs are associated with all of these procedures for preserving data confidentiality when storing and processing sensitive personal information from longitudinal studies.

At the data analysis, external dissemination, and reuse stages

The defining features of longitudinal data also raise confidentiality challenges at the data analysis, and external dissemination, and reuse stages of the research information lifecycle. Researchers using or reusing longitudinal data for analysis generally prefer to have access to the data in microdata form, a form that allows linkage with other datasets but also leaves personal information more vulnerable to disclosure risks. At the same time, traditional statistical disclosure techniques are often inappropriate for creating deidentified longitudinal data files. Consequently, researchers may rely on licensing agreements and restricted access for disseminating data as alternatives to releasing public-use longitudinal data.

Researchers may seek to disseminate their longitudinal datasets for a variety of reasons. The sponsors of their research may mandate the release of survey data in open or privacy-preserving forms where possible.[45] In addition, scholarly journals often have policies that expect or mandate the sharing data along with publication.[46] More generally, the norms of scientific communities generally encourage the release of open data for the purposes of replication by other researchers.[47] Although these policies encourage researchers to disseminate their data widely, a number of factors constrain their ability to do so.

Most longitudinal studies store the raw information collected from subjects in microdata files

[45] *See, e.g.*, National Institutes of Health, "Data Sharing Policy and Implementation Guidance," last updated March 5, 2003, http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm; Bureau of Labor Statistics, "National Longitudinal Surveys: Frequently Asked Questions," last accessed October 22, 2013, http://www.bls.gov/nls/nlsfaqs.htm.
[46] For a discussion of journal data policies in greater detail, see the workshop paper covering the Access to Journal Publication Data use case, which will be available at http://privacytools.seas.harvard.edu/fall-2013-workshop.
[47] *See id.*

containing linked records. Unlike aggregate or summarized data, microdata contain information at the level of the individual human subject. Longitudinal microdata enable researchers to answer complex questions, test alternative hypotheses, and calculate marginal effects of changes over time.[48] Data in this form also permit identification of data anomalies or errors, and support replication of results by third party researchers.[49] Because longitudinal studies collect data at level of the individual for the purposes of studying patterns in individual behavior, data that have been aggregated or summarized, such as through the use of cross-tabulation tables, are often not suitable for certain types of longitudinal analysis. In particular, aggregated longitudinal data are ill-suited for analyses not anticipated by the researcher who produced the aggregate dataset. The often sensitive nature of the data, the potential to link the data to data from other sources, and the need to analyze and share the data as microdata create challenges for sharing longitudinal data with other researchers or disseminating it to the public.

In many cases, the heightened disclosure risks make the dissemination of raw, identifiable longitudinal data impossible, and researchers may, instead, manage disclosure risks in the data by applying statistical disclosure limitation (SDL) techniques to created deidentified data files. SDL techniques, which alter and redact data values, are commonly used to transpose raw datasets into public-use or restricted-use datasets that limit confidentiality risks and can be released to the public. However, as discussed below in section 5, traditional SDL techniques developed to address disclosure risks in cross-sectional data are often not appropriate for managing confidentiality in longitudinal data. While traditional SDL techniques may reduce risks associated with privacy and confidentiality, they often do so at the expense of data utility, and the tradeoff between privacy and utility is heightened for longitudinal data. A successful SDL treatment should preserve the important analytical properties of the data, but, when applied to longitudinal data, it is likely to change the structure of the data in ways that sharply influence the statistical models and inferences made by a secondary researcher. As a result, certain types of modeling or analysis might not be possible. Compounding this problem is that the types of SDL treatments that are applied to a dataset are not always disclosed to the public. Secondary researchers sometimes unknowingly treat a deidentified release as a raw data release, which may affect the results of their analyses.

In light of the challenges associated with other approaches, researchers often use licensing agreements and remote and limited access to manage access and usage rights when sharing longitudinal data. Restricted access regimes in conjunction with data repositories or enclaves are often used to manage access rights and conditions for sharing raw microdata with project collaborators and secondary researchers. For example, researchers may release a deidentified public-use dataset, but also make raw data available under a restricted access regime.

When developing a restricted access regime for sharing longitudinal data, researchers take a number of factors into consideration. Researchers require the ability to analyze the data, potentially from geographically diverse locations, without exposing the data to confidentiality or

[48] *See* Judith D. Singer and John B. Willet, *Applied Longitudinal Data Analysis* (New York: Oxford University Press, 2003).
[49] *Id*.

security risks. Longitudinal studies typically span extended periods of time, so restricted access mechanisms must be sufficiently robust to adapt to personnel changes (including changes to principal investigators, researcher assistants, contractors, secondary researchers, and others associated with a study) over periods of many years or decades. Data enclaves can be very costly to establish and, therefore, are appropriate for large research projects or data releases that can be managed through existing data enclaves. Access restrictions will also vary depending on the categories of users accessing the records. For example, commercial entities, such as marketers, will be treated differently from academic researchers. Alternatively, access may be restricted to agency employees, or researchers working under federal grants, to ensure that all data users will be subject to federal confidentiality and security regulations. These factors require a flexible means for disseminating or providing access to the data to primary or secondary researchers, and potentially the public at large, that preserves confidentiality.

In addition to technical controls, data use agreements are often used as a form of legal protection, particularly in cases where data are shared with external collaborators and secondary researchers. Data use agreements have been adopted by federal and state government agencies, academic institutions, data repositories, data enclaves, research studies, nonprofit organizations, and commercial entities. Such agreements describe the contents and sensitivity of the data; the restrictions on access, use, and disclosure; the data provider's rights and responsibilities; the data confidentiality, security, and retention procedures to be followed; the assignment of liability between the parties; and relevant enforcement procedures and penalties. The parties involved in a transfer of research data utilize such an agreement in order to set forth obligations, ascribe liability and other responsibilities, and provide a means of recourse if a violation occurs.

4 Risks associated with longitudinal data

Many of the disclosure risks associated with storing any type of data are heightened with longitudinal data. The risk of data loss, which may result from the loss of a contractor's laptop or a breach by an attacker, increases in a longitudinal study. One factor contributing to the elevated risk is the fact that a large number of contractors and field investigators, often spread across a wide geographic area and over a long period of time, are provided with the names, addresses, and phone numbers of study participants in order to re-contact them during each wave of a study.

In addition to generalized risks, such as data loss and confidentiality breaches, longitudinal data are subject to other risks born of their defining characteristics. In a longitudinal study, a vast amount of likely sensitive, individual-level data are often collected from subjects over an extended period of time and released in periodic intervals over the course of the longitudinal study. As more data are collected, stored, and shared over the course of a longitudinal study, the risks associated with the data, such as the potential for linkage to external data sources, necessarily increase. In addition, the vulnerability of sensitive, highly-specific microdata and the elevated likelihood of discovering a subject's participation in the study increase the disclosure risks associated with longitudinal data.

Although the value of longitudinal studies lies in part in their ability to link a set of behaviors and changes to each individual over time, this characteristic tends to make the combination of

observations associated with each subject, or quasi-identifiers,[50] unique and thus potentially identifiable. The disclosure risk in a longitudinal study is higher due to the cumulative and highly-specific nature of the individual-level data collected and stored. Observations of change across several waves of a longitudinal study are likely to generate unique combinations, such as a location change paired with an occupation change, that would provide strong indicators of an individual's identity.

Longitudinal data may also be associated with particularly high disclosure risks depending on the level of sensitivity of the personal information collected. The information is collected, stored, and shared as data at the individual level and, in many cases, contains answers to highly-sensitive questions that cover topics such as an individual's sexual and criminal activities. These factors increase the risk that an individual could be identified within the dataset and that disclosure would cause significant harm. For these reasons, the level of sensitivity of the data is a factor related to disclosure risks at the collection, storage, use, and sharing stages of the research data lifecycle, and data administrators typically adjust their data security procedures depending on how sensitive a given dataset is.

Another risk associated with longitudinal data is that outside individuals might discover that a subject participated in the study. One criterion used for assessing the potential disclosure risk of a dataset is that an attacker who has no knowledge about who has taken part in a study should not be able to identify an individual. However, if an attacker knows a particular individual is in a survey sample, he or she can identify the individual's record with greater certainty. A relative, neighbor, or friend may learn that an individual participated in a longitudinal study if a field investigator contacts such people in the process of locating a participant, the subject herself reveals her participation status, or a study involves cluster sampling and includes many neighbors, colleagues, and students from the same region or establishment. Under any of these scenarios, there is a heightened risk that the participant may be more easily reidentified within the research dataset. Sometimes it is not possible to eliminate the risk of others learning about a respondent's participation in the study, such as in the case of longitudinal studies that interview each member of a household or every student in a school. Also, repeated contacts between investigators and participants increase the likelihood that an individual's participation will be discovered.

Although the disclosure risks associated with longitudinal data may be higher than those associated with other types of data, the frequency of reidentification and the extent of disclosure harms are relatively unknown. The privacy harms resulting from a disclosure—often described in general terms such as possible loss of employment or insurability; criminal liability; emotional, reputation, or dignitary harm; or price discrimination—are not very well understood. In addition, the legal remedies availability to victims are limited.[51] Reidentification in general is a relatively new, rarely-reported, and little-understood threat. It is not clear who all of the potential attackers

[50] A quasi-identifier is a set of attributes that in combination can uniquely identify individuals (e.g., birth date and gender). *See* Latanya Sweeney, "*k*-Anonymity: A Model for Protecting Privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10 (2002): 557, 564.
[51] *See* Felix Wu, "Privacy and Utility in Data Sets," *University of Colorado Law Review* 84 (Fall 2013): 1117; Ryan Calo, "The Boundaries of Privacy Harm," *Indiana Law Journal* 86 (Summer 2011): 1131.

are, how disclosed longitudinal data would be monetized or otherwise used by an attacker, or if a reidentification incident would affect the researchers' ability to continue the study.

The legal system is evolving to address some of these concerns. Contractual approaches to data sharing and use generally contain terms assigning liability and penalties in the event of a breach. Federal and state legislatures frequently debate proposed changes to existing statutes and regulations. For instance, many state laws and some federal statutes have established notice duties in the event of data breaches. This legislative activity has helped to raise public awareness of data breaches in the commercial sector. However, the ability to recover damages through a lawsuit remains limited due to the burden of showing that an actual harm has occurred as a result of a breach. As a matter of jurisprudence, many courts are reluctant to award damages in cases where the injury is merely an increase in the *risk* that a future harm might occur. The harm must be one that the law recognizes, such as a financial or substantial emotional injury. In many cases, it can be difficult for a victim to prove that a disclosure incident directly caused a particular harm.

5 Current approaches to managing confidentiality in longitudinal data

To address some of the challenges and risks associated with longitudinal data, researchers employ a combination of legal and technical methods. Contractual approaches such as consent agreements, institutional policies, and data use agreements are among the most common legal techniques for managing confidential longitudinal data. Technical tools, including an array of available statistical disclosure limitation techniques and data access controls, are also commonly used to protect security and confidentiality in longitudinal data.[52]

Legal instruments play a significant role in managing data confidentiality, mostly in the areas of consent, access, use, and oversight. Privacy statutes and institutional policies, including institutional review board policies, typically contain provisions requiring researchers to obtain the consent of human subjects participating in a study and to limit access to and use of confidential research data. Accordingly, data use agreements that prescribe access to and use of a given dataset are very common, but oversight and enforcement continues to be a difficult problem. Common contractual approaches to enforcement include sanctions such as denial of future access to data files, reporting of violations to federal agencies or a researcher's institution or funders, and fines and other statutory penalties. There have been recent proposals in this area, including a proposed legislative and contractual framework that would create a safe harbor for transfers of data for research uses and establish rights and remedies for data subjects as third party beneficiaries of data use agreements.[53]

Prior to collecting data, researchers generally must obtain informed consent from participants. For respondents age seventeen and younger, interviewers also request consent from the respondents'

[52] For a general overview of statistical disclosure limitation techniques, *see* Federal Committee on Statistical Methodology, *Report on Statistical Disclosure Limitation Methodology*, Statistical Policy Working Paper 22 (December 2005), http://www.fcsm.gov/working-papers/SPWP22_rev.pdf; *see also* Anco Hundepool, et al., *Statistical Disclosure Control* (West Sussex, UK: Wiley, 2012).
[53] *See* Robert Gellman, "The Deidentification Dilemma: A Legislative and Contractual Proposal," *Fordham Intellectual Property, Media & Entertainment Law Journal* 21 (2010): 33.

parents or guardians. Because laws and regulations, the scope of data collection, and privacy expectations change over time, researchers generally pay special attention to updating consent agreements for repeated collections of data. As a result, a participant's consent may be seen as valid for only a particular data collection, and, each time a respondent is asked to participate in a new collection activity, her consent is again required. This means there are regular opportunities to obtain respondents' consent as circumstances change and the study evolves over time.

It is common for data holders to establish a restricted access regime for maintaining confidential research data, particularly when dealing with highly sensitive personal information. Data holders may limit data access to a certain class of researchers, such as members of a university community or researchers working under a federal pledge of confidentiality, or require secondary researchers to submit applications requesting access to the data. They often also require secondary users to agree to participate in confidentiality training or adopt a data security plan to safeguard the data. Alternatively, secondary researchers may be provided with access to the data only through remote or physical data enclaves.

Researchers have reported that data holders sometimes err on the side of being too restrictive in allowing access to their data. Many holders of longitudinal data rely on some form of restricted access, not because it is not possible to release data more widely while still protecting subject confidentiality, but, rather, because they lack the tools to assess the threats and risks of their datasets.[54] Some data holders attempt to produce public-use datasets, but the primary mechanism for release seems to be restricted access. This is an especially common practice for the sharing of longitudinal data. Because it is difficult to anticipate the risks that will be associated with the release of future waves of a study, researchers are often inclined to choose more restrictive access controls when sharing data from longitudinal studies. Many researchers report that restricted access is an imperfect solution but that they lack better tools for preserving data confidentiality. If there were a technical tool that would achieve the same confidence in privacy and confidentiality as restricted access regimes do, many practitioners would prefer to make data available using such a tool.

Some other technical tools and techniques, such as statistical disclosure limitation methods, are used to create datasets that can be shared more widely, or even with the general public. Broadly speaking, the goal of these techniques is to disguise and obscure data or limit the degree of interaction with raw data. Traditional SDL methods include aggregating, suppressing, and perturbing data. Aggregation involves rounding and topcoding certain values[55] to make them less precise; suppression entails removing some of the most sensitive data from a dataset before sharing it with others; and perturbing means altering some of the data, such as by introducing noise or by swapping some of the values. When preparing data for public release, researchers often suppress information that directly or indirectly identifies a research subject and perturb the data by swapping certain values among similar respondents or by introducing random noise to the

[54] Lawrence H. Cox, Alan F. Karr, and Satkartar K. Kinney, "Risk-Utility Paradigms for Statistical Disclosure Limitation: How to Think, But Not How to Act," *International Statistical Review* 79, no. 2 (2011): 160-99, 179, http://onlinelibrary.wiley.com/doi/10.1111/j.1751-5823.2011.00140.x/pdf.
[55] *See, e.g.*, Bureau of Labor Statistics, "National Longitudinal Surveys: Frequently Asked Questions," last accessed October 22, 2013, http://www.bls.gov/nls/nlsfaqs.htm.

data. In practice, multiple disclosure limitation techniques are often used in combination to protect the confidentiality of a dataset.

There is wide variation among researchers regarding their knowledge and adoption of sophisticated SDL techniques. Government agencies, such as the US Census Bureau, have published SDL guidelines that influence how other organizations design their data releases. Even where robust SDL techniques have been adopted, statisticians have noted the difficulty of using traditional SDL techniques to preserve the confidentiality of longitudinal data. Models for assessing the disclosure risk in a given dataset are typically developed using cross-sectional data and are poorly suited for addressing longitudinal risks.[56] Likewise, the SDL techniques that are effective for cross-sectional datasets often result in either weaker privacy protections or a reduction in data utility when applied to longitudinal data.[57]

As discussed above, long-term longitudinal datasets typically contain vast quantities of highly-specific, and perhaps even uniquely-identifying, data about individuals. The combination of these pieces of information collected over a long period of time, including the links and patterns inherent to longitudinal data, can reveal sensitive and uniquely-identifying information about individuals.[58] As a result, there is a heightened risk that an adversary could link a longitudinal dataset to other datasets, including public records, in a way that could compromise the privacy of individuals.[59] In addition, even when researchers attempt to reduce disclosure risks by releasing aggregate data using cell suppression techniques, the suppressed cells may be vulnerable when the release contains data from multiple scales (such as data compiled at both the county and state level) and from multiple data collections over time.[60] Given the heightened disclosure risks associated with longitudinal data, applying traditional SDL methods to such data in a way that provides robust privacy protections will often lead to a substantial loss in utility. For these reasons, there are not many examples of successful applications of SDL techniques to longitudinal data in order to produce public-use datasets.

More modern technical approaches, such as differential privacy and synthetic data, are less widely utilized to protect the confidentiality of longitudinal data, though they may better preserve the complex relationships between variables in a longitudinal dataset. By releasing synthetic data,

---

[56] *See* Lawrence H. Cox, Alan F. Karr, and Satkartar K. Kinney, *supra* note 54.

[57] *See generally* Khaled El Emam and Luk Arbuckle, "Longitudinal Discharge Abstract Data: State Inpatient Databases," chap. 4 in *Anonymizing Health Data: Case Studies and Methods to Get You Started* (Sebastopol, CA: O'Reilly Media, 2013); Benjamin C.M. Fung, et al., *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques* (New York: CRC Press, 2010).

[58] *See generally* Lawrence H. Cox and Laura V. Zayatz, "An Agenda for Research in Statistical Disclosure Limitation," *Journal of Official Statistics* 11, no. 2 (1995): 205-220, http://www.census.gov/srd/papers/pdf/lvz9301.pdf; National Research Council Committee on National Statistics, *Improving Access to and Confidentiality of Research Data: Report of a Workshop* (2000), http://www.nap.edu/catalog.php?record_id=9958.

[59] See, e.g., Stephen E. Fienberg, "Privacy and Confidentiality in an e-Commerce World: Data Mining, Data Warehousing, Matching and Disclosure Limitation," *Statistical Science* 21, no. 2 (2006): 143-154, http://arxiv.org/pdf/math/0609288.pdf.

[60] *See* Scott H. Holan, et al., "Bayesian Multiscale Multiple Imputation with Implications for Data Confidentiality," *Journal of the American Statistical Association* 105, no. 490 (2010): 564-77, http://www.bls.gov/osmr/pdf/st080010.pdf.

or simulated microdata, researchers may be able to reduce disclosure risks while retaining validity for certain inferences that are consistent with the model used for synthesis.[61] One example of a synthetic longitudinal dataset is the public release synthetic Longitudinal Business Database, released in 2011 by the US Census Bureau and the Internal Revenue Service as an experimental product to be improved over time.[62] Researchers have used the differential privacy criterion to characterize the confidentiality protection in a small portion of the synthetic Longitudinal Business Database, but this is very much a nascent area of research.[63]

Workshop participants expressed a high degree of interest in the potential of differential privacy to overcome many problems associated with preserving the confidentiality of longitudinal data. So far, though, there are very few differentially private algorithmic results that apply to the setting that is typical to longitudinal studies. More precisely, differential privacy has mostly been studied in the "one-shot" setting: a dataset is gathered once and for all, and information about it is then released, either as a single publication or interactively in response to queries from users. Less attention has been devoted to datasets that are built up over time and analyzed "online," as the data are acquired. The most common model along those lines is the "continual observation" model.[64] Although the model is flexible enough to describe longitudinal studies, existing work (to our knowledge) assumes that one person's information affects only a limited number of stages of the study.[65] For instance, the literature does not directly address the design of methods that release accurate statistics about each stage of a study as well as detailed information about how the statistics are evolving over time.

6 Conclusions

Key issues related to the confidentiality of long-term longitudinal data

The defining characteristics of a long-term longitudinal study—that such a study follows a large number of individuals over extended periods of time and collects vast quantities of sometimes very sensitive personal information from subjects—create a number of challenges for researchers working with longitudinal data. Over the course of a longitudinal study, data collection, uses of data, privacy expectations, and laws may change, leaving the scope of subjects' consent unclear. The use of contractors for data collection and the need to retain administrative data linked to records to recontact subjects over time may increase the risk of disclosure.

---

[61] Ashwin Machanavajjhala, et al., "Privacy: Theory Meets Practice on the Map," *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering* (2008): 277-86, http://www.cse.psu.edu/~dkifer/papers/PrivacyOnTheMap.pdf.

[62] *See* Satkartar K. Kinney, et al., "Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database," *International Statistical Review* 79, no. 3 (December 2011), http://hdl.handle.net/10.1111/j.1751-5823.2011.00153.x.

[63] *See id.*

[64] *See* Cynthia Dwork, et al., "Differential Privacy under Continual Observation," *Proceedings of the ACM Symposium on the Theory of Computing* (2010).

[65] *See, e.g.,* T.-H. Hubert Chan, Elaine Shi, and Dawn Song, "Private and Continual Release of Statistics," *ACM Transactions on Information and System Security* 14, no. 3 (2011): 26; Prateek Jain, Pravesh Kothari, and Abhradeep Thakurta, "Differentially Private Online Learning," *Proceedings of the 25th Conference on Learning Theory* (2011): 1-34.

Laws and regulations require researchers to disclose to participants the extent of confidentiality protections and the scope of research to be conducted in longitudinal studies. An institutional review board may oversee the research design process and the confidentiality-preserving techniques used by the primary researchers. A variety of technical and legal mechanisms are commonly used to safeguard the confidentiality of longitudinal microdata. Current approaches to managing confidentiality in longitudinal data include using legal instruments, such as data use agreements, which set forth the permissible conditions under which secondary researchers may access, use, and share the data with others, and technical tools, such as statistical disclosure limitation methods like aggregation, suppression, and perturbation, to protect confidentiality in longitudinal data releases. Data enclaves and archives, in combination with restricted access and restricted data regimes, also play a role in preserving confidentiality by providing a technical infrastructure for managing access privileges and the confidentiality of longitudinal data over the course of its lifecycle.

Urgent problems and areas for future research

Workshop participants flagged a number of urgent problems related to confidentiality in longitudinal research data. The following is a discussion of some of the problems identified and areas where additional research is most needed.

- *Law-technology interaction.* There is a need for researchers to understand better the interaction between law and technology in the context of managing confidentiality in longitudinal data. Researchers in this area have not identified where the law works best, where technology works best, and how best to combine legal and technical methods where a blended approach is most appropriate. The current approach of most practitioners is to apply stringent legal restrictions and disclosure limitation methods, but using these tools independently of each other often has the effect of unnecessarily reducing data availability and utility. Future research should focus on developing tools and best practices that enable the legal and technical tools to work together, with the goal of making longitudinal data more widely available while preserving both utility and privacy.

- *Access to advanced technical solutions.* Recent developments in advanced technical solutions such as differential privacy are promising, but such techniques have not yet been widely implemented. There is a particularly high demand for advanced technical solutions for the sharing of longitudinal data because researchers have encountered significant difficulties in applying traditional statistical disclosure limitation techniques to longitudinal data. At the same time, there are also challenges associated with applying differential privacy techniques to longitudinal data. Future research should be directed to developing implementations of these tools, identifying where they work best and where new solutions might be helpful, and educating data holders and researchers about their use. Such solutions should be made widely available to researchers at reasonable cost.

- *Shared infrastructure.* It is important to develop shared infrastructure for managing confidential data. Infrastructure such as data enclaves and remote statistical analysis tools can be very costly, and, in most cases, it is not necessary to clone such infrastructure in

each institution. Shared infrastructure is especially critical to the widespread implementation of advanced techniques like differential privacy.

- *Data ownership, breach liability, and other enforcement mechanisms.* Data ownership is not very well defined. The operational definition is access control, and contractual approaches to breaches focus on liability shifting between parties sharing data. Though contracts provide a means of legal recourse, they are often difficult to enforce at scale. Making significant changes to privacy statutes and regulations, such as by instituting strict liability or stronger criminal liability provisions, may be a viable solution, but it is not clear such changes are desirable. Before reshaping the law in this area, the relevant stakeholders must decide where rights, responsibility, ownership, and liability for research data should reside.

- *Outreach and education.* The development of best practices and the widespread adoption of more advanced disclosure limitation tools depend on researchers becoming aware of the utility and privacy benefits of these new techniques. In addition, public perception of data security and confidentiality plays an important role in the viability of longitudinal studies. The public's view of whether privacy is being adequately protected when collecting, storing, using, and sharing vast amounts of personal data influences potential subjects' willingness to participate in a research study. As more privacy-preserving tools are developed and implemented, practitioners should also invest in outreach and education materials that explain the benefits of these new tools to researchers, potential study participants, and the general public.

- *Cost-benefit analysis.* Public dialogue about the future of longitudinal research and confidentiality will undoubtedly have a normative component. Researchers should evaluate the costs and benefits of longitudinal data from a societal perspective and be prepared to discuss how the cost-benefit analysis will change over time with the adoption of new legal and technical approaches to data confidentiality.

Additional research in these areas will be critical to advancing the integration of various technical and legal methods for preserving confidentiality in long-term longitudinal data.